CGIAR Research Program on
Climate Change, Agriculture and Food Security (CCAFS)

Example Data Management Plan at Activity Level

October 2013

These materials were produced for and with funding from the Climate Change Agriculture and Food Security Research Program of the Consultative Group on International Agricultural Research (CGIAR).

Introduction

This plan describes the data management activities related to the CCAFS Household Baseline Survey.

The basic data management principles related to the CCAFS programme are:

- Data generated by the project will be placed into the public domain with restricted access for confidential data such as names and GPS coordinates;
- All data generated by the project are owned by the CGIAR consortium;
- Informed consent must be obtained from respondents who must be made aware that they can withdraw from the study at any time.

Data Collection

Capture Methods

The aim of this activity is to gain an understanding of current practices among farmers and how these have changed in recent years. To this end we will be conducting an interviewer-led household survey.

Paper copies of the questionnaires will be completed by the interviewers. Questionnaires will be translated into the local language prior to the interviews.

Data Description

The survey will initially be run in 15 sites spread over 12 countries from the three regions of West Africa, East Africa and South Asia. In each site a block with interesting agricultural systems and institutional links will be selected and we will list all the villages within each block. From this list we will randomly select seven villages. In each of the seven villages we will list all the households and from these lists randomly select 20 households from each village to be surveyed.

Thus, we expect to have 140 households for each site. Each site will have a unique identifier which will be a 4-character code – the code will consist of 2 letters representing the country plus 2 digits. Blocks will be numbered uniquely within the survey. Each selected village will have a unique code which is 4-characters and can be any combination of letters and digits. Within the villages the selected households will be numbered. Thus, the unique identifier for a household will be the combination of Site ID, Block ID, Village ID and Household ID.

We will collect data on the following:

- The most important crops currently grown;
- Whether this has changed in the last 10 years;
- The most important types of livestock kept by the households and whether this has changed in the last 10 years;
- What farming practices have changed in the last 10 years and reasons for these changes;
- Do these changes affect specific crops and/or livestock;
- What items are currently produced on-farm and what are produced/collected off-farm;
- What access do farmers have to land and water;

- Do they use fertiliser;
- Do they have access to information regarding climate events; e.g. information regarding the start of the rains or long and/or short-term forecasts;

We will also be asking questions regarding:

- Food security – e.g. are there times in the year when they struggle to find enough to eat;
- Other sources of income (besides farming);
- Membership of community groups;
- Assets owned by the household;
- General demographic information such as household size and composition, education level and family type (male or female-headed).

For tracking purposes, we will collect names of the household head and the main respondent – we expect these to be the same in most cases – together with GPS coordinates of the dwelling. This tracking information will not be put into the public domain and is solely for enabling us to revisit the same households in the future.

Most of the variables will be coded and a code sheet will be produced with a set of standard codes. We expect additional crop and livestock codes to be used in each site and after data collection the data manager will work on merging these crop and livestock codes to produce a comprehensive consolidated list. This will include appropriate recoding in the data files. A document will be produced detailing the merging method used and list any recoding that will have been done to the data.

To record the amount of land that households have access to we will record and use the local unit of measurement. The supervisor will then record the conversion factor to convert these values to hectares.

Computerisation & Storage

Data Entry

Data will be recorded on paper questionnaires which will be visually checked for completeness and consistency by both the enumerators and the supervisors. Both will sign and date the front of the questionnaire when they have done these checks. A checklist will be produced to ensure that the same checks are done on each completed questionnaire.

For data entry we will be using CSPro. A consultant will produce the data entry system with screens that resemble the questionnaire as far as possible. The data entry system will be programmed to follow the skips present in the questionnaire. A data entry manual will be produced to accompany the system, and this will be used when training the data entry staff. Once the questionnaire is finalised a list will be drawn up of possible consistency and range checks that can be programmed into the data entry system. The documentation will detail these checks.

A version of the data entry system with the screen labels in French will be produced for use in West Africa.

The data entry system will be thoroughly checked by entering data from 5 completed questionnaires and adjustments made as appropriate.

Data entry staff will receive 2 or 3 days of training on using the system. They will each receive a copy of the manual and a log book to record problems.

Quality Assurance

At the end of each day the field supervisors will visually check each completed questionnaire for obvious errors. Instructions for these checks will be included in the Field Supervisors Manual.

Double data entry will be used. For each site two data entry clerks will each enter data from all 140 questionnaires into separate data files. The data entry supervisor or data manager will then use the CSPro Data Compare tool to look for differences in these files. Any differences will be checked against the original questionnaires and corrected in both data files. The output report from the Data Compare tool will be stored with the project archive as part of the audit trail of data quality checks.

After the double data entry comparisons further checks on the data will be done. Frequency tables will be produced on all coded variables to ensure that there are no values outside the expected range. Frequency tables on the villages will help us check that we have the correct number of households per village. Names and GPS coordinates will be checked against the sampling frame files to make sure we have collected data from the sampled households.

For the land values we will do the following:

- Check that the total amount of land owned is always greater than the amount of owned land used for grazing, growing crops, etc.
- Do the same for rented land – i.e. the amount of rented land used for grazing cannot be greater than the total amount of rented land.
- Look at the highest and lowest values to ensure they are reasonable and that no one farmer has an order of magnitude more land than any other farmer. We will produce histograms or boxplots of land values so that we can easily see the range of data. Any extreme values will be investigated.
- The data will be compared against local knowledge – for example, we might know that in a particular site, farmers tend to have very small plots of land, so if the data are showing very large farms then we will investigate. We need to ensure for example that the conversion factor is recorded correctly. The conversion factor would be the number of land units in a hectare.

A log will be kept of potential errors and outliers in the data, together with a report of how we dealt with these values.

Once the data are checked and corrected a data quality assessment document will be produced.

Data Structure & Organisation

The main study unit will be the household and we expect 140 household level records per site. There may also be data at other levels – for example plot level, individual level, etc. Until the questionnaire is finalised we will not know what other levels there will be. The data entry system will be set up so that data from all levels will be entered at the same time into the same CSPro file.

Once exported, data from each level will be stored in separate data files and each will include the primary key fields from the household level, namely Site ID, Block ID, Village ID and Household ID. These fields will act as the link between data at different levels.

The data will be entered using CSPro but once the data comparisons and data checks are completed, they will be transferred to SPSS. There will be one CSPro data file but separate SPSS files for each level in the data. Syntax files will be produced for labelling the data and for calculating some standard indices. As yet the set of standard indices have not been decided but these will be documented when ready.

Data Dictionary

A data dictionary will be produced containing the following information for each variable:

- **Name**– variable names will be no more than 8 characters in length and the names will be printed on the final questionnaire so that researchers can easily see which question a variable refers to.
- **Label –** the label will be the question text or an abbreviation of it.
- **Codes and labels –** any numeric codes used will be listed. A code book will be produced for cases where several variables use the same set of codes and the dictionary for the variables will refer to the code book.
- **Missing value codes** – we will use three missing value codes throughout
    - **-8** will be used to indicate "Not applicable"
    - **-6** will indicate no consensus among family members
    - **-9** will be for any other missing value
- **Unit of measurement** – the unit of measurement for amounts of land will vary between sites; therefore, there will be a variable containing the name of the local unit and a second variable containing the conversion factor to convert the local units to hectares.

The primary key will be the combination of Site ID, Block ID, Village ID and Household ID. The set of derived variables has not yet been decided.

Storage & Sharing

The data manager will be responsible for producing guidelines for project members on the structure of the shared folders – the DDS. The guidelines will include how to name files and the expected times and modes of delivery for project files.

We will be using Dropbox to share project files among team members. One Dropbox share will have read/write access to all team members but there will be a separate share to which only the data manager has write access and others are included by way of links. In this share the data manager will start to build the archive and will inform team members of any updates and additions via email. All files in the read-only share will include dates in the file name – dates will be in the form YYYYMMDD – e.g. HouseholdQuestionnaire20110516.docx

Files on Dropbox will be backed up automatically, but the data manager will also make daily backups of the Dropbox folders. These will be stored on the DMs own PC.

Legal Aspects

Ethics & Privacy

An information sheet will be prepared for respondents which describes the study. This, and the consent form, will clearly state that the data collected will be used for research purposes and that data will be made public although any identifying information will be removed prior to archiving. The box below shows the consent statement that will appear at the top of the questionnaire.

> *"Good morning/afternoon. We are coming from (_partner organisation's name_) with permission from the local government. We are conducting a survey* ***looking at farming practices and how they change over time****. We would like to ask you some questions that should take no more than one to one and half hours of your time. We would like to share some of this information widely in order that more people understand how food is grown and used in this region and the issues that you face regarding food production and soil, water and land management.*
>
> *Your name will not appear in any data that are made publicly available. The information you provide will be used purely for research purposes; your answers will not affect any benefits or subsidies you may receive now or in the future. Do you consent to be part of this study? You may withdraw from the study at any time and if there are questions that you would prefer not to answer then we respect your right not to answer them.*

Prior to archiving, names and GPS coordinates will be removed from the main data files. Only village codes will be used in the data files rather than village names. Separate files containing the names will have restricted access in the archive.

Data Ownership

All team members are aware of the data ownership agreements of the project and have signed the agreement. Any new team members will be asked to sign the agreement before being able to access the data.

Copyrighted Material

Digital maps used are copyrighted.

Archiving & Preservation

A Dataverse will be created for the Baseline studies and data and documentation will be uploaded. The original CSPro data files will be loaded into the Dataverse but will have restricted access. The archive will include final versions of all files from the household survey up to and including site reports. The following files will be included:

- Questionnaires – in all languages
- Code book
- Fieldworker manual
- Data entry manual
- Data checking guide
- Data entry system
- CSPro data files – restricted access
- Analysis plan

- SPSS syntax files for labelling data, calculating indices and carrying out standard analyses
- Output from analysis plan
- Merged and anonymised data files in SPSS format
- Household Identification information – restricted access
- Sampling Frames – restricted access
- Site analysis reports
- Process reports – e.g. data quality assessment document

The Dataverse will be set up within 24 months of the end of data collection.

Training & Responsibilities

Within each country team an individual will be given data management responsibilities and will report to the overall project data manager. The data manager will have overall responsibility for ensuring the files are submitted in a timely manner.

Data entry staff will receive training on using the data entry system. Data comparisons will be done by the local team if there is anyone in the team with the required skills. Otherwise the files will be sent to the project data manager for the comparisons to be done.